Assessing diagnostic and screening tests: Part 2. How to use the research literature on diagnosis

Ruth Gilbert
Center for Evidence Based Child Health
Department of epidemiology and statistics
Institute of Child Health
London

Stuart Logan
Systematic Reviews Training Unit
Institute of Child Health
London

Virginia A Moyer
Center for Clinical Research and Evidence Based Medicine
Department of pediatrics
University of Texas-Houston Health Science Center

Elizabeth J Elliott
Department of paediatrics and child health
University of Sydney and Children's Hospital at Westmeade
Sydney, Australia

Correspondence to:
Dr Moyer
virginia.a.moyer@uth.tmc.edu

Competing interests:
None declared

Case scenario

One night when on call for inpatient admissions, you receive the third call in a week to admit a child who looks well but has fever and a petechial rash. Few of these children actually seem to have bacterial (particularly meningococcal) sepsis, but you are not sure how good the "diagnostic test" of the child's clinical appearance (well vs not well) performs in this circumstance. You estimate that it would be 20 times worse to miss a child who has bacteremia than to admit a child unnecessarily (that is, the action threshold is around 5%) and wonder whether the child's appearance can be incorporated into the decision to admit children who have fever and petechial rash. You frame the question, "In children with fever and petechiae (population), does looking ill (diagnostic test) increase the risk of bacteremia (outcome)?" and search for literature on this topic using the search strategy "petechiae AND bacteremia" in the "Clinical Queries" section of PubMed's web site (highlighting *diagnosis* and *sensitivity*). This strategy yields 24 studies.

Summary points

- In appraising the research on a diagnostic test, physicians should ask whether the results are likely to be valid (close to the truth) and whether they are applicable to the physician's clinical setting
- A test is useful only when an action threshold could be crossed based on the result of the test
- Precise pretest and post-test probabilities are often difficult to calculate, but a few simple calculations can provide an upper and lower estimate of the probability of disease
- Knowing the threshold for taking action, the pretest probability of disease in a patient, and the likelihood ratio for the test will help in the interpretation and application of the test results

SEARCHING THE LITERATURE FOR THE BEST AVAILABLE STUDY

For questions about medical interventions, sources such as the *Cochrane Library, Best Evidence,* and *Clinical Evidence* contain predigested evidence, in which someone else has gone to the trouble of performing a thorough search, has appraised the evidence, and has synthesized the valid studies for many clinical questions. Unfortunately, for questions about diagnosis, few such sources of predigested evidence exist. Searching the original literature is easiest if some methodologic terms are incorporated into the search strategy; the "Clinical Queries" feature of PubMed actually does this by incorporating a previously tested set of strategies to maximize either the sensitivity or the specificity of the search. A sensitive search will yield the most citations, but many may be irrelevant; a specific search will yield fewer, but some relevant publications may be missed. If the Clinical Queries search is unsuccessful, proceed to a search without methodologic terms on PubMed or any other search engine for MEDLINE.

After 1 or more articles are found that address the question, the next step is to decide whether the result can be believed. This involves critically appraising the study methods to decide whether the results are likely to be valid (close to the truth) and whether the results (if the study is likely to be valid) are applicable to a specific setting. The following guides (see box) will help physicians decide whether the study they are reading is valid and applicable to their patients.

IS THE STUDY VALID?

Does the study include an independent, blind comparison with an adequate reference standard?

Ideally, a reference standard represents unequivocal truth, and with its use, it should be clear which patients definitely have the disease and which definitely do not. For example, chromosome analysis for trisomy 21 is the reference standard for Down syndrome. However, for most conditions, the division into those who have the disease and those who do not is an artificial cutoff in a spectrum of disease severity. For example, the reference standard for anemia in a toddler is generally quoted as a hemoglobin level of less than 11 g/L (<110 g/dL). But why is it not 11.5 or 9.0 or 10.5 g/L? A cutoff has to be drawn somewhere to define which patients require treatment or further investigation and which do not. Although few reference standards are perfect, physicians need to judge whether the reference standard used in an article is acceptable. This will depend on how closely the reference standard relates to prognosis and the potential to benefit from intervention.

The next step is to work out whether there was a blind comparison between the test and the reference standard. It is important that knowledge of the result of the "test" did not influence the decision to obtain the reference standard and that knowledge of the reference standard did not influence the result of the test. Unblinded comparisons are biased toward agreement. If a colleague says, "Listen to this heart; I think there is a systolic murmur," other physicians are likely to agree. Similarly, if the lead physician (for instance, of a team) says, "This heart sounds normal," others would be less likely to hear a murmur. Tests are independent if the reference standard does not include the test or elements of the test. For example, the white blood cell count (test) should not be compared with a reference standard for sepsis that is composed of blood culture, white blood cell count, and clinical condition because here the reference standard includes the test being evaluated.

Did the study sample include an appropriate spectrum of patients to whom the test would be applied in practice?

Physicians need to work out whether a test's use was compared in all patients or just in those who obviously had the disease and those who obviously did not. The latter often

John Radcliffe Hospital/SPL

In children with fever and petechiae, looking ill increases the likelihood of bacteremia

happens. Clinicians decide to write an article, pull out a selection of their obviously "bad cases," and do the reference standard in a group of unaffected "controls." Common sense dictates that it is easy to differentiate between severe disease and no disease at all. Any reports like this should be treated with caution because they exaggerate test performance.[1] Moreover, such studies provide no information on patients at intermediate risk of disease—a group most in need of effective testing. Ideally, the physician should look for a study that has evaluated the test in patients with a similar range of disease severity as his or her own patients. It is rarely possible to find the perfect study. Studies from referral centers will often be found, and the physician will have to judge how their results apply to the patients at hand. A rough rule is that patients with negative results are less likely to be referred, so the proportion of false-negatives is underestimated (and sensitivity is overestimated). The excess of patients with positive results leads to an overestimation of the proportion of false-positives (thereby underestimating specificity). Similarly, studies in primary care, where few patients have comorbid conditions that give rise to false-positive results, are likely to overestimate specificity. In practice, the spectrum of patients in which the test was evaluated, characterized by disease severity, comorbidity, or the age or sex of patients, usually does affect the likelihood ratio (LR).[2-4] The LR should, therefore, be thought of as an average measure of test performance.

Did the test result influence the decision to perform the reference standard?

Gold standard tests may be expensive, invasive, or hazardous, so usually a test is needed to replace the reference standard; however, when evaluating the test, the reference standard should be performed regardless of the test results. For example, if a physician wants to know about the performance of prenatal serum testing for babies with Down syndrome, he or she would search for a study in which all women had an amniocentesis and in which the serum test result did not affect management. If women knew the results of their tests, those with a negative result might not turn up for amniocentesis, and the proportion of false-negative results would be underestimated. Sometimes it is simply not feasible to perform the reference standard on all patients with a negative result. In these cases, the reference standard can be performed on a random sample of patients with a negative result. More information on the pitfalls and calculations associated with this approach has been reported elsewhere.[5] In some studies, different reference standards are used for patients with positive and negative test results. Such studies can overestimate test performance.[1]

In the search described in the case scenario, 1 article was found[6] that directly addresses the question posed in the scenario. The reference standard in this study was blood culture, probably the best reference standard available for bacteremia. The person interpreting the blood culture did not know whether the child looked ill, and the person judging the child's clinical appearance did not know the results of the blood culture, so a blind comparison is likely. The appearance of the child likely did not influence whether a blood culture was done; virtually all children with fever and petechiae (393/411) had a blood culture result available. The patient spectrum appeared to be broad: the authors meticulously surveyed all 24,000 patients with a temperature of 38°C or higher (≥100.4°F) who attended an accident and emergency department during an 18-month period. They identified 411 patients with petechiae, 8 of whom had bacteremia. The patients studied are, therefore, likely to be similar to the hypothetical patient population in the scenario. Given the answers to the questions so far, the study appears to be reasonably valid.

Critical appraisal criteria for studies of diagnostic tests

Is the study valid?

- Does the study include an independent, blind comparison with an adequate "gold" or reference standard for the diagnosis?
- Did the study sample include an appropriate spectrum of patients to whom the test would be applied in practice?
- Did the test result influence the decision to perform the reference standard?

What are the results?

- What is the likelihood ratio?
- How precise is the likelihood ratio?

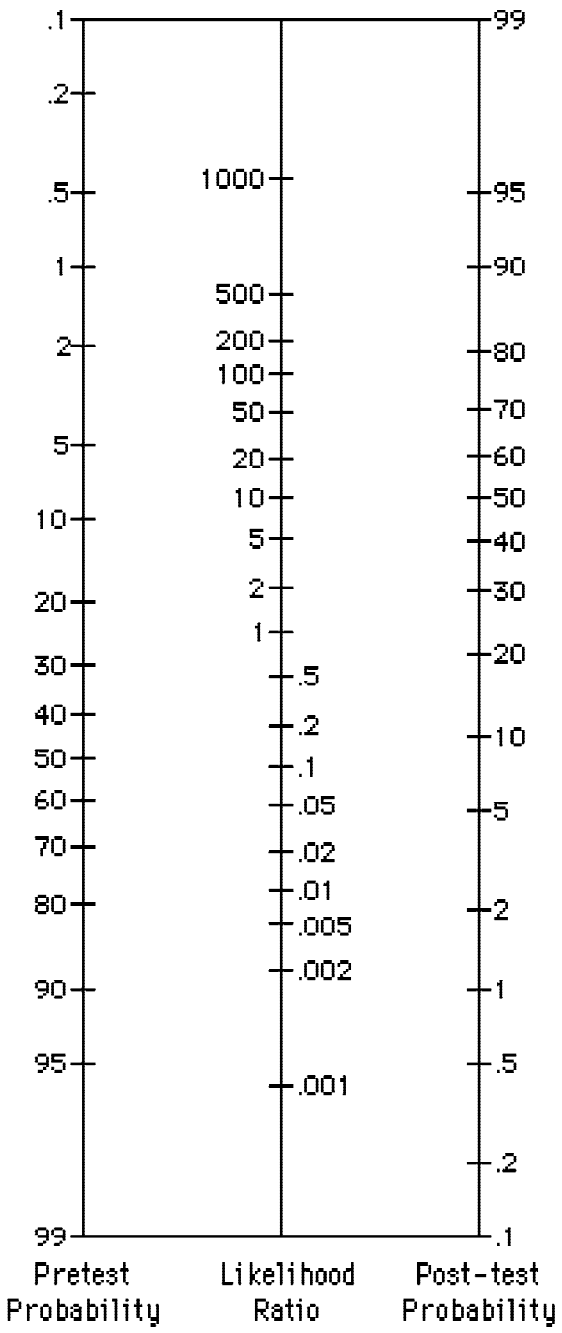
Will the test results help with patient care?

- What is the pretest probability of disease in a specific patient population?
- What is the post-test probability of disease, and does it cross the action threshold?
- Were the methods of performing the test described in sufficient detail to be reproducible in practice? Is the test feasible and affordable, accurate, and precise in other settings?

WHAT ARE THE RESULTS?

What is the LR?

Calculation of the LR was discussed in the previous article in this series.[7] Briefly, the LR compares the proportion of those with the disease who have a particular test result with the proportion of those without the disease who have the same test result. If the proportions are the same, then the LR is 1.0, and the test does not distinguish between those



The likelihood ratio nomogram (adapted from Fagan[8])

with and those without the disease. The further the LR is from 1.0 (higher or lower), the better it distinguishes one group from the other.

How precise is the LR?

The LR is an estimate of how well the test works, based on the sample of patients in the study, so it is subject to variation between studies. If the sample size is small or the number of patients in a group is small, then the estimate may be off, and the confidence interval (CI) is used to quantify that impression. The 95% CI is used to describe the range within which the true result is likely to lie 95% of the time. A wide 95% CI indicates an estimate that is not precise, and a narrow one indicates a precise estimate of the LR. If the 95% CI is not given, keep in mind that the smaller the study, the wider the CI will be, and the less precise the results will be.

WILL THE TEST HELP WITH PATIENT CARE?

What is the pretest probability of disease in a specific patient population?

From clinical experience, from local data about a specific population, or from published studies of the probability of various diseases, the clinician should be able to estimate the probability that a patient will have the disease before a test is performed. The estimate need not be exact. A range of possibilities can be set that seem reasonable, and then the LR applied to each of them to determine whether, if any of them were true, the action threshold would be crossed.

What is the post-test probability of disease, and does it cross the action threshold?

Having named a pretest probability, or a range of probabilities that reasonably apply to a specific patient or patient population, the LR nomogram (figure)[8] is used to determine the post-test probability. A straight edge is laid across the pretest probability on the right-hand column and the LR (for the test result that was obtained) on the center column. The post-test probability can now be read on the right-hand column. If the pretest and the post-test probabilities would lead to different actions, then the action threshold has been crossed. There is no point in obtaining a test whose results would not cause a change in management (whether it causes a physician to take action or to choose not to take action). A test is useful when an action threshold could be crossed based on the result of the test.

Were the methods for performing the test described in sufficient detail to be reproducible in practice? Is the test feasible and affordable, and is it accurate and precise in specific settings?

The reproducibility and precision of any subjective assessment are important, particularly if the study was performed by 1 skilled researcher rather than by people who would be likely to use it in practice. If, in the study being appraised, a single skilled researcher performed the test, then test performance will likely be better than in routine practice. For example, ultrasonography of nuchal thickness as a diagnostic test for Down syndrome may be

highly reliable in the hands of a team of fetal medicine specialists (the same result occurs in the same patient time and again, by the same or a different observer) but less reliable in the hands of nonspecialized ultrasonographers. If a study is based on a small number of testers, information on the intraobserver and interobserver variation is also needed.[9]

QUESTION THE GOLD STANDARD

Much of this article has been based on the assumption that we know what the reference standard means. However, few reference standards unequivocally distinguish disease from no disease, and often the reference standard is itself an arbitrary cutoff in a spectrum of disease. For example, there is no clear dividing line between children who have cerebral palsy and those who do not. Most clinicians will agree about a child with severe spastic quadriplegia, but how would children with mild monoplegia or moderate dyspraxia be characterized?

For some tests (blood pressure, fasting blood glucose, postnatal depression), the test result must be correlated with the patient's eventual prognosis and response to treatment, and the physician must decide at what level of severity the benefits of intervention outweigh the harms. Evidence about how test results relate to benefits of treatment can be obtained from randomized controlled trials (RCTs). If RCTs are not available or do not relate to an appropriate patient group, cohort studies may provide useful information. Evidence relating to prognosis can be obtained from cohort studies and sometimes from controlled trials, although the populations in most controlled trials are too highly selected to be representative of the population at large.

RESOLUTION OF THE SCENARIO

Mandl et al systematically evaluated 24,000 children with fever, but only 8 were found to have petechiae and bacteremia, which highlights the difficulty of performing such studies.[6] Likelihood ratios are not reported in their article, but they can be calculated by completing a 2×2 table for "appears ill" or "does not appear ill" compared with the reference standard (bacteremia). The LR for looking ill (that is, a positive test result) was 6.4, with a 95% CI of 3.9 to 10.4, and the LR for does not look ill (that is, a negative test result) was 0.28, with a 95% CI of 0.08 to 0.94. The patients in Mandl et al's study are similar to yours, so the pretest probability of bacteremia in your patient is likely to be similar to that of the patients in their study, which was 2% (8/410). The post-test probability is read from the nomogram by joining the pretest probability (2%) and the LR for looking ill (6.4), giving a post-test probability of about 12%. The post-test probability of 12% is well above your action threshold of 5% for admission. A "sensitivity analysis" in which the lower 95% CI for the LR of 3.9 is used produces a post-test probability of 7.9%, still well above your action threshold. If the child did not look ill, the LR is 0.28, so the post-test probability would be about 0.8%. Even the higher end of the 95% CI, 1.9%, would be below the threshold for admission. This test result appears to have the potential to change a clinical action, so it is worth obtaining.

Can this test be applied in practice? Clearly, any one physician's assessment of an ill child may differ from the assessment used in the study. Fortunately, the authors described in some detail what they meant by "ill," so the use of their criteria in practice is straightforward.

CONCLUSIONS

Diagnosis may be complicated and involve a series of tests or observations. Precise pretest and post-test probabilities are often difficult to calculate, but the use of these principles can provide an upper and lower estimate of the probability of disease. The guidelines discussed here can aid in deciding which studies of diagnostic tests are likely to be valid and applicable to a particular group of patients. Finally, interpretation and application of diagnostic tests will be made easier and more effective (cause most benefit for least harm) if the physician is more explicit about the threshold for action, the pretest probability of disease in patients, and the LR for the test.

References

1 Lijmer JG, Mol BW, Heisterkamp S, et al. Empirical evidence of design-related bias in studies of diagnostic tests. *JAMA* 1999;282:1061-1066 [erratum published in *JAMA* 2000;283:1963].
2 Bossuyt PM. No burial for Bayes' rule [editorial]. *Epidemiology* 1997;8:4-5.
3 Ransohoff DF, Feinstein AR. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. *N Engl J Med* 1978;299:926-930.
4 Brenner H, Gefeller O. Variation of sensitivity, specificity, likelihood ratios and predictive values with disease prevalence. *Stat Med* 1997;16:981-991.
5 Begg CB. Biases in the assessment of diagnostic tests. *Stat Med* 1987;6:411-423.
6 Mandl KD, Stack AM, Fleisher GR. Incidence of bacteremia in infants and children with fever and petechiae. *J Pediatr* 1997;131:398-404.
7 Gilbert R, Logan S, Moyer VA, Elliot EJ. Assessing diagnostic and screening tests. Part 1. The concepts. *West J Med* 2001;174:405-409.
8 Fagan TJ. Nomogram for Bayes' theorem. *N Engl J Med* 1975;293:257.
9 Pham B, Klassen R. Assessing clinical measures and clinical disagreement. In: Moyer V, et al, eds. *Evidence Based Pediatrics and Child Health.* London: BMJ Books; 2000:71-78.